Chemometrics Modelling of Environmental Data

Roma Tauler, Emma Peré-Trepat, Silvia Lacorte and Damià Barceló

Department of Environmental Chemistry, Institute of Chemical and Environmental Research, IIQAB-CSIC, Jordi Girona, 18, 08034, Barcelona, Spain, e-mail rtaqam@iiqab.csic.es

Abstract: Environmental monitoring studies produce huge amounts of concentration values of chemicals spread at distant geographical sites and during different time periods. Moreover, the content of chemicals is also estimated at different environmental compartments (i.e. air, water, sediments, biota...). All these data values are difficult to cope and evaluate in a simple and fast way using simple univariate statistical tools, specially due to their large number and to their multivariate correlation. In order to discover relevant patterns within large multivariate data sets, the application of modern chemometric methods based in statistical multivariate data analysis and in Factor Analysis is proposed. The basic assumption of chemometric methods is that each of the measured parameter in a particular sample is affected by contributions coming from multiple independent sources. Each one of these sources is characterized by a particular chemical composition and is distributed among samples in an unknown way. After applying chemometric methods, point and diffuse sources of contaminants in the environment and their origin (natural, anthropogenic, industrial, agricultural...) are identified and their relative distribution among samples (geographical, temporal, among environmental compartments) evaluated. At each sampling site, relative source quantitative apportionment is estimated allowing a global evaluation of the environmental impact, distribution and evolution of main chemical contamination sources in the environment. In this presentation, different chemometric methods will be tested on a series of environmental data sets. In particular, the application of principal component analysis and multivariate resolution methods is shown to be a powerful tool for the goal of chemometrics modelling of contamination sources in large environmental data sets acquired in monitoring studies.

Keywords: Modelling, Chemometrics, Principal Component Analysis, Multivariate Curve Resolution

1. INTRODUCTION

Chemometric data analysis methods (Massart et al., 1998) provide powerful tools for the analysis and interpretation of large, environmental, multivariate data sets generated within environmental monitoring programs (Einax et al. 1997). The goal of these studies is the computation, screening and graphical display of patterns in large data sets, looking for possible groupings and sources of data variation. The basic assumption of these multivariate exploratory data studies is that main sources of data variance observed in the concentration changes of contaminants are due to a reduced number of contamination sources of different origin (industrial, agricultural,...) defined by profiles describing their chemical composition and their geographical and temporal distributions. Large environmental analytical data sets containing concentration information of multiple chemical compounds collected at different sampling sites and at different sampling periods are arranged in large tables, data matrices, or in more complex data structures according to different dimensions, modes, orders or directions of experimental measurement (Zeng Y et al. 1990). In the chemometrics literature, these complex data structures are commonly called multiway data sets or higher order tensor data sets (Geladi 1989, Smilde 1992).

Principal Component Analysis (PCA, Joliffe 1986, Wold et al. 1987) is one of these multivariate statistical methods frequently used in exploratory data analysis. PCA allows the transformation and visualization of complex data sets into a new perspective in which the more relevant information is made more obvious. Using PCA, contamination sources may be identified and their geographical and temporal distributions estimated. A complementary approach proposed to achieve similar results is Multivariate Curve Resolution using Alternating Least Squares (MCR-ALS, Tauler 1995). Whereas PCA is intended mostly for identification and interpretation of contamination sources, MCR-ALS is proposed for the resolution

of the 'true' underlying contamination sources. In this work, these and other multiway data analysis approaches based in PARAFAC and Tucker models (Henrion, 1994) will be proposed and compared for the analysis of large environmental monitoring data sets. Both approaches, PCA and MCR-ALS, are extended to the analysis and interpretation of multiway data sets obtained in exhaustive monitoring programs.

Summarizing, the main objective of this work is to show how Principal Component Analysis, Multivariate Curve Resolution and other multiway data analysis methods can be applied in the investigation of environmental data sets from exhaustive monitoring studies in order to: a) identify and interpret the main contamination sources present in a particular data set; and b) determine their geographical, temporal and among compartments distributions

2. ENVIRONMENTAL DATA TABLES (Figure 1)

Environmental data sets are usually organized in data tables or data matrices, corresponding to one sampling time period or environmental compartment of the monitoring campaign, giving K data matrix arrays of I rows corresponding to I (geographical) sampling sites and J columns corresponding to J measured variables (concentrations of chemical contaminants or other environmental parameters). Variables having very few values above the detection limit should be removed before multivariate data analysis is applied. When a particular compound is not detected, its concentration value is set equal to half its detection limit (Fharnham, 2002). For missing values, imputation methods have been proposed (Walczak, 2001) and whenever they are a small fraction of the measured values, they may be estimated without loosing the data structure needed for application of multivariate and multiway data analysis tools. Statistical descriptive plotting methods like box plots provide useful tools for data overview, fast visual data variance examination and outliers' description. However they do not allow the description and interpretation of multivariate relationships nor the detection, interpretation and resolution of the underlying (latent) multicomponent sources of data variation.

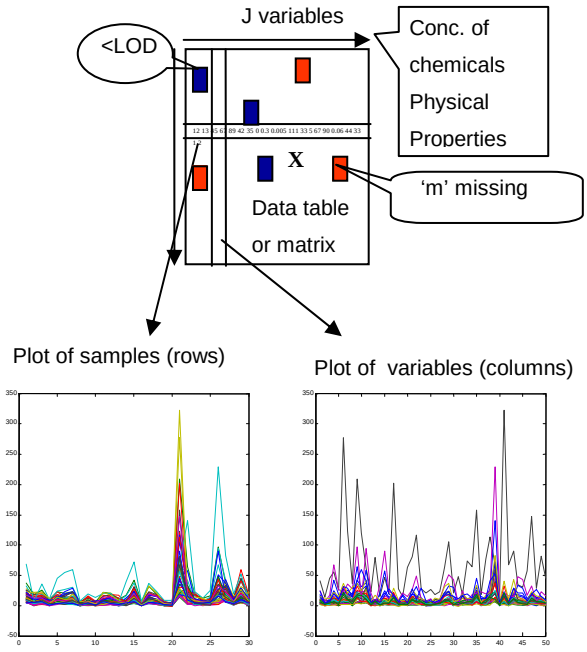


Figure 1. Environmental Data Tables

3. CHEMOMETRIC MODELS AND METHODS

Data pretreatment methods usually employed in chemometric data analysis studies include mean centering, scaling, autoscaling and log transformation. Mean centering removes constant background contributions, which usually are of no interest for data variance interpretation. However, mean centering may produce undesired effects if resolution and apportionment of 'true' environmental sources is intended, since it gives negative values. On the other hand, in some environmental compartments like surface waters, mean centering has little effect on the results since most of the values of the different variables are so low that their average is also very low and close to zero. Some kind of data scaling is mandatory when variables are of different type and their values are at different scales and units. Scaling to unit variance has a notorious variance effect since it increases the weight of variables that initially have lower variances and decreases the weight of those which have higher initial values and variances. In some cases, this effect may distort significantly the results of data analysis making interpretation more difficult, especially for these variables having only very few values larger than the detection limit. When the same errors are expected for all the measurements of one variable, column norm scaling is an adequate way to give similar weight to

all different measured variables. Log transformation of experimental is also recommended for skewed data sets, like those in environmental studies where the majority of the values are low values with a minor contribution of high values. With log data pretreatment, a more symmetrical distribution of experimental data is obtained; however, a loss of the internal linear data structure may occur and more linear components are needed to explain the same amount of data variance. In order to remove negative values from input data before log calculation, a constant value, usually equal to 1, is added to all the entries, or even better, values are changed of scale (e.g. from mg/kg to μg/kg) In this way, log values resulted to be non-negative. Finally, tables of binary correlations between pairs of variables may be also easily calculated and evaluated.

To investigate multivariate correlations, identify and interpret multicomponent contamination sources and deduce their geographical, temporal and among environmental compartment distributions, Principal Component Analysis, PCA method and Multivariate Curve Resolution Alternating Least Squares, MCR-ALS (Tauler, 1995) are proposed. Both approaches assume a linear model to explain the observed data variance using a reduced number of components:

$$x_{ij} = \sum_{n=1}^{N} g_{in} f_{nj} + e_{ij} \qquad \text{Equation 1}$$

$$X = G F^{T} + E \qquad \text{Equation 2}$$

In equation 1, x_{ij} refers to the measured concentrations of chemical component j in sample i, f_{nj} refers to the contribution of variable j (chemical compound j) to the environmental source n, and g_{in} refers to the contribution of source contribution n to sample i. e_{ij} gives the unexplained contribution considering the total number of n=N environmental sources. This equation means that the measured concentrations are a weighed (scores, g_{in}) sum of a reduced number (N) of main environmental contributions defined by a particular chemical composition (loadings, f_{nj}), apart from noise (multiple small unknown contributions) and experimental error defined by e_{ij}. The weights or scores g_{in}, describe how the main contamination sources are distributed among the analyzed samples and the loadings f_{nj}, identify the chemical composition of these contamination sources. When this linear equation is written in matrix form (equation 2), X is the matrix of measurements, G is the matrix of scores (distribution of contamination sources among samples), F is the matrix of loadings (composition of the composition sources) and E is the noise or error matrix containing the variance not explained by the model defined by the N environmental sources described in G and F .

Both PCA and MCR-ALS methods are based on this bilinear model. Since only X is initially known, matrix decomposition described by equation 2 is not unique (ambiguous) unless constraints are applied. PCA constraints F and G solutions to be orthogonal. F moreover is also normalized and forced to be in the direction of explaining maximum variance. Components (loadings and scores in F and G) are extracted in a stepwise way, i.e. the first component explaining maximum variance, the second component explaining the remaining maximum variance, once first component contribution has been subtracted, and so on. Under such constraints, PCA provides unique solutions and interpretation of variance is straightforward since scores and loadings are orthogonal (not overlapped). Using a small number of principal components a considerable amount of data variance is usually explained since many of the analyzed variables are correlated. Therefore, interpretation and visualization of main features and trends of the data set under study, i.e. of main contamination sources, are readily available from score and loading plots. However, this PCA decomposition does not estimate the 'true' underlying (latent) sources of data variance but a linear combination of them fulfilling orthogonal constraints. Scores and loadings evaluated by PCA apart from orthogonal can be negative. This means that although these solutions have good mathematical properties, they do not have a physical meaning (chemical concentrations and geographical or temporal distributions never can be negative)

A possible complementary and/or alternative method to perform the matrix decomposition given in equation 2 is MCR-ALS (Tauler, 1995). In this case, loadings and scores are not constrained to be orthogonal like in PCA, but to fulfil a particular set of physical constraints like non-negativity (non-negativity alternating least squares optimization). The goal of such a decomposition is to recover how contamination sources are really in physical terms (loadings) and how do they really are really distributed among samples (scores). However, since only matrix D is known and only soft constraints like non-negativity and normalization are applied, unique solutions are not guaranteed and rotational and intensity ambiguities may be present (Tauler 1995).

The bilinear model shown in equation 2 may be easily extended to the simultaneous analysis of multiple data sets using data matrix augmentation Thus, bilinear methods like PCA and MCR-ALS are easily adapted to three-way and multiway data sets (Tauler 1995) by matrix augmentation or cube unfolding (matricizing). More complex trilinear and multilinear models preserving the data structure have been proposed also for the investigation of environmental contamination sources. In particular trilinear models for three-way data are described by the two equations:

$$x_{ijk} = \sum_{n=1}^{N} g_{in} f_{jn} z_{kn} + e_{ijk} \qquad \text{Equation 3}$$

$$X_k = G Z_k F^T + E \qquad \text{Equation 4}$$

In equation 3, x_{ijk} are the measured concentrations of chemical component j at sample i under condition k. There are three ways, orders or modes of measurement. These three modes indicate that component j was analyzed at sample i at a particular situation or condition k, usually time or environmental compartment (water, sediment or biota). The whole data set can be organized in a data 'cube' or parallelepiped as shown in Figure 2.

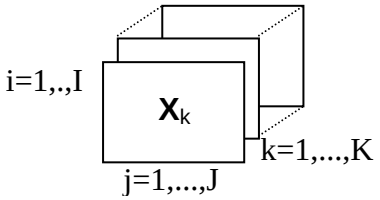


Figure 2. Three-way data arrangement

where $\mathbf{X_k}$ is the slice or matrix k of the data parallelepiped, which is modelled by equation 4, where Z_k is a diagonal matrix. This trilinear model described by equation 4 is also called the PARAFAC model (R.Bro, 1997). In the trilinear model, all slices in the three-way data set are decomposed using the same G (scores) and F^T (loadings), differing only in their relative amounts expressed in the different Z_k diagonal matrices. Trilinear models, and by extension multilinear models, provide unique decompositions and they are the natural extension of bilinear models. They are useful for data exploration and interpretation. However, since they impose equal scores and loading profiles for all data matrices simultaneously analyzed, they are in many circumstances, too rigid, and do not allow the resolution of the 'true' underlying sources of data variation, simply because the data do not behave like in the postulated trilinear models. A compromise between 'softer' bilinear models and 'harder' trilinear models should be considered in practice according to the data structure encountered for a particular data set.

RESULTS AND DISCUSSION

In order to know if different chemometrics methods work satisfactorily and to evaluate pretreatment and rotational ambiguity effects, different two-way and three-way data sets have been simulated fulfilling respectively a bilinear and/or a trilinear model:

Case 1. Two-way bilinear data

Case 2. Three-way trilinear data

Case 3. Three-way non-trilinear (bilinear) data

Factor loadings and scores are simulated assuming log distribution of values and pseudo-random proportional error contributions. Effect of pretreatment methods for different data structures are evaluated by singular value decomposition and principal component analysis. In general, scaling and log transformation increase the relative contribution of minor components and they may be recommended depending on the case.

PCA gives scores and loadings more difficult to interpret than MCR-ALS, which provides simpler factor profiles, practically equal to those used for the data simulation. See for instance results in Figure 3. The agreement between MCR-ALS resolved first loading (red) and the actual loading used for the simulation (blue) is excellent. The same happen with other factor loadings and scores used in the simulation. In the case of the analysis of simulated three-way data, application of methods based on trilinear models give only an accurate factor resolution if data are strictly trilinear, failing in cases where data deviate from this ideal situation (Figure 4). Correlation coefficients between 'true' and PARAFAC resolved profiles (see Figure 4) are not good enough.

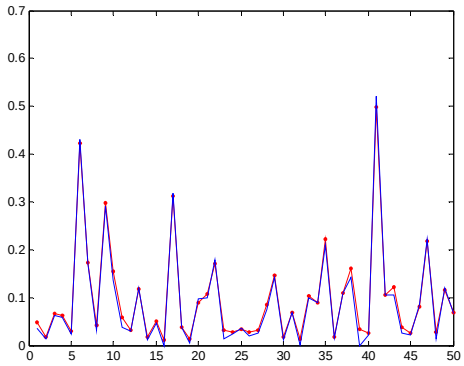


Figure 3. Comparison 1st loading 'true' (blue) vs 'mcr-als' (red)

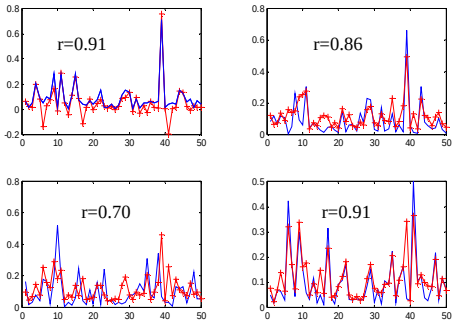
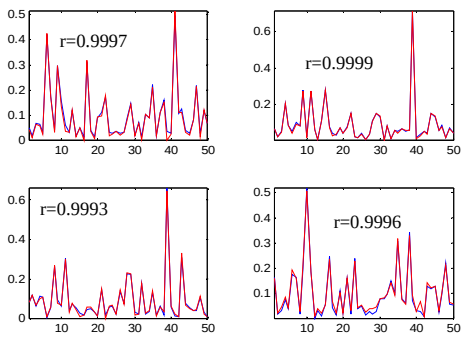


Figure 4. PARAFAC results for non-trilinear data. Comparison of resolved loadings 'true' (blue) vs 'PARAFACs' (red)

Conversely, in the case of MCR-ALS without assuming a trilinear model, an optimal resolution and fit of the experimental data is achieved and correlation coefficients between 'true' and MCR-ALS resolved profiles are very good (all r > 0.999) for all of them. In practice this will be a common situation in the analysis of complex environmental multiway data sets, where the higher flexibility of bilinear models allow a better resolution and fit of the experimental data. This is also a situation frequently encountered for many chemical data sets (Tauler, 1995).



5. REFERENCES

Figure 5. MCR-ALS results for non-trilinear data. Comparison of resolved loadings 'true' (blue) vs 'mcr-als' (red)

Results obtained in the analysis of a large experimental data set obtained in an exhaustive study of contamination sources of semivolatile organic compounds used as herbicides in surface river waters of Portugal (Tauler et al. 2004) were confirmed by the results obtained in this work concerning the study of simulated data. Main contamination sources of semi volatile organic compounds in surface waters of Portugal were identified and resolved by application of different chemometric methods. These contamination sources had different origins: agricultural, for simazine, atrazine, alachlor, and metholachlor in central and south of Portugal; industrial, specially for tributhylphosphate in the Porto and Ave River areas (north of Portugal); and mixed for 4-chloro-2-methylphenoxy)acetic acid, 2,4-dichloro-phenoxy) acetic acid and mecoprop widespread used in the whole Portugal geography. Temporal distribution profiles of these contamination sources in the one-year period covered by this study showed peak values in spring and summer seasons. Deeper conclusions about geographical distribution and temporal evolution of these contamination sources would require a more extensive analysis of data acquired in multiyear monitoring programs. Similar interpretations about the more important contamination sources (loadings) and about their geographical and temporal distribution (scores) were possible using different chemometric methods, increasing the reliability of the conclusions achieved in this work. The proposed method for averaging PCA and MCR-ALS unfolded score profiles resulted to be an efficient and useful way to uncover mixed geographical and temporal information from two-way bilinear models when applied to three-way data. In this way also, information obtained by these methods can be easily compared with the information provided by score profiles obtained using three-way methods like PARAFAC. See Tauler et al. (2004) for more details about this work.

Bro, R., PARAFAC. Tutorial and a applications, *Chemom Intell Lab Syst.*, 38, 148-171, 1997.

Einax, J.W., H.W. Zwanzinger and S. Geiss, *Chemometrics in Environmental Chemistr,*. VCH, Weinham, Germany, 1997.

Fharnham, I.M., A.K Singh, K..J. .Stetzenbach and K.H. Lohannesson, Treatment of non-detects in multivariate analysis of groundwater geochemistry data, *Chemom Intell Lab Syst*, 60, 265-81, 2002.

Geladi, P., Analysis of multiway (multimode) data. *Chemom Intell Lab Syst.*, 7, 11-30, 1989.

Henrion, R., N-way principal component analysis theory, algorithms and applications, *Chemom Intell Lab Syst* 25, 1-23, 1994.

Joliffe, I.T., *Principal Component Analysis*,. Springer, New York, NY, USA, 1986..

Massart, D.L., B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*, Elsevier Science, Amsterdam, Holland, 1998.

Smilde, A.K., Three-way analysis. Problems and projects, *Chemom Intell Lab Syst*, 15, 143-157, 1992.

Tauler, R., A.K. Smilde and B.R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *J Chemom*, 9:31-58, 1995.

Tauler, R.,. Multivariate curve resolution applied to second order data, *Chemom Intell Lab Syst*, 30, 133-146, 1995

Tauler, R., S. Lacorte, M. Guillamón, R. Cespedes, P. Viana and D. Barceló, Resolution of main environmental contamination sources of semivolatile organic compounds in surface waters of Portugal using chemometric compounds., *Environmental Toxicology and Chemistry*, 23, 565-575, 2004

Walczak, B., and D.L. Massart, Dealing with missing data, Parts I and II. *Chemom Intell Lab Syst*., 58, 15-27 and 29-42, 2001.

Wold, S,, K Esbensen and P. Geladi, Principal component analysis. *Chemom Intell Lab Syst* 2,37-52, 1987.

Zeng, Y. and P. Hopke, Methodological study applying three-mode factor analysis to three-way chemical data sets. *Chemom Intel Lab Sys,t* 7, 237, 1990.